DOCUMENT RESUME

ED 260 459 CS 209 290

AUTHOR McDaniel, Barbara A.
TITLE Ratings Vs. Equity in the Evaluation of Writing.
PUB DATE Mar 85
NOTE 24p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (36th, Minneapolis, MN, March 21-23, 1985).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Chinese; Comparative Analysis; *English (Second Language); Higher Education; Holistic Evaluation; Interrater Reliability; Scores; Test Theory; Test Validity; *Writing Evaluation; Writing Instruction; *Writing Research
IDENTIFIERS Primary Trait Scoring

ABSTRACT

A study was conducted to determine whether evaluators of large scale essay tests respond the same way toward essays written by English as a second language (ESL) and non-ESL students. The data examined came from the English Placement Test (EPT) administered in the province of British Columbia, Canada, in March 1979. The test was used to identify three types of students: those who needed no remediation, those who had difficulty with English because it was not their native language, and those who were native speakers of English but who had severe language or writing problems. The placement test had two parts: an objective part containing 40 error recognition sentences, and a subjective portion that required the writing of an essay. Evaluators of the essays were to score them analytically in three separate categories: (1) content development and organization, (2) sentences, and (3) words. Two thousand papers out of the total 18,000 were randomly selected for the sample. Native English speakers were labeled "E-1 writers" and ESL students (among whom Chinese was the dominant first language) were labeled "E-2 writers." Results suggested that evaluators of large scale essay tests do not appear to respond in the same way toward essays written by ESL and non-ESL students and that researchers will have to define more specifically how to judge what is "good" and "bad" and how much weight to give these things. (Tables of findings are included and the English Placement Test Notes to Markers is appended). (DF)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED260459

# RATINGS VS EQUITY IN THE EVALUATION OF WRITING

by

Barbara A. McDaniel

Department of English and Communications
Douglas College
New Westminster, British Columbia, Canada

March 1985

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Barbara McDaniel
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



CS 209 290

Ratings VS Equity in the Evaluation of Writing

The December 1982 issue of College Composition and Communication devoted half its pages to a major aspect of writing instruction: testing students' proficiency. Half a dozen articles composed by people at the forefront of knowledge on testing discussed procedures and issues related to large-scale tests. The following month, January, 1983, the second issue of Notes From the National Testing Network in Writing appeared. This publication from the City University of New York and the Fund for the Improvement of Post-secondary Education also contained many good articles and short notes on a wide range of topics in testing. And both publications contained a statement from the Chief Investigator of the Conference on College Composition and Communication Task Force on Testing, which called for further research on testing "to protect the equity and integrity of the educational process."[1] But despite considerable concern expressed regarding the reliability and validity of large-scale tests of writing proficiency, the articles mentioned above say little about scoring. Obviously, assessment cannot be reliable and valid unless raters' results are consistent and they are scoring what they are assumed to be scoring.

It is difficult to appraise such things as the perceptions and the consistency of raters, yet the goal of "equal access to educational opportunities"[2] for all students requires that scoring get the same kind of attention given the other stages of test design.

Not only has too little been written about scoring large-scale tests; scarcely any research examines the appropriateness of using the same tests or the same evaluators for native and non-native speakers of English.

Native and non-native speakers have learned English in different ways, and members of the two groups can have mastered different things, i.e., can differ in their strengths and weaknesses in ways that can escape the notice of persons not trained to perceive them. The purpose of the pilot study discussed below was to see whether people who mark large-scale essay tests appear to respond in the same way toward sentences written by the two groups, or whether, lacking special training, markers were inconsistent in how they scored the two.

I will use the phrase "native-English writers" or "E-1 writers" for the group under discussion who spoke English as their first language. I will use "ESL writers" or "E-2 writers" for those who spoke another language before they spoke English and who have signs of interference from that language in their written English (ESL is widely used to stand for English-as-a-second-language).

Lest the mistake be made of thinking the questions raised here concern only specialists in English as a second language, several facts should be known. Almost anyone who teaches English these days encounters or should expect to encounter ESL students. The numbers of people in North America who have acquired English as an additional language have grown rapidly in the past decade. In the mid-seventies the United States already had an estimated 28 million persons--one in eight--with non-English language backgrounds.[3] But since then, worldwide political and economic oppression has so increased immigration in the United States, much of it illegal, that both liberal and conservative politicians say "We have lost control of our borders."[4] The U.S. Congress is trying to "stop the flood"[5] of immigrants, but bills being considered may give amnesty to many illegal

residents, thus increasing the pressure on schools to provide equitable education for language minority groups.

Another reason ESL problems should concern all educators is that non-native English speakers no longer tend to live in specific areas. In 1976-77 it was estimated that the language-minority population of 23 states had reached ten percent of the total population, or more; in three states, the proportion was one in four. Finally, Canada has a unique concern because of her officially bi-lingual status. The Canadian commitment to equality of opportunity for English and French speakers, which was reaffirmed in the 70's, made all the more relevant, in dominantly English provinces, the necessity of evaluating the English of native French speakers equitably.

Examining the fate of ESL writers on large-scale tests is useful because raters of such tests reflect general attitudes toward writing. In a manner of speaking, they represent teachers in general, i.e., their perceptions indicate what is happening in classrooms everywhere. Problems in evaluation suggest broader problems in teaching. It is, therefore, important that teachers' perceptions be understood when equity is an issue.

## I. Background of the Study

The data examined here came from the English Placement Test (EPT) administered in the Province of British Columbia, Canada, in March of 1979. The EPT had been established in 1976 because several post-secondary institutions had asked the provincial government to help them with the placement of incoming students. In general, the institutions who used the placement test had to identify three types of students: those who needed no "remediation" of any kind, "remediation" being work in English grammar,

usage, and composition intended to bring students' writing to standards the post-secondary schools expected in written work; those who had difficulty with English because it was not their native language; and those who were native speakers of English or the equivalent but had severe language or writing problems. The placement of ESL students was the most serious concern because of the large number of such students in British Columbia public schools; roughly fifty percent of the students in Vancouver, British Columbia's largest city and the third largest city in Canada, are ESL students. As citizens or legal immigrants in Canada, many of these seek entrance to the public colleges and universities. The March 1979 EPT was administered to all graduating high school students in British Columbia and to a small number of people who had finished school but needed the scores for admission to certain British Columbia institutions. This research was conducted at Simon Fraser University in British Columbia, Canada.

The placement test had two parts: an objective portion containing forty error-recognition sentences, and a subjective portion, requiring the writing of an essay. Directions for writing the essay called for a composition from three to five hundred words in length on the writer's choice of three topics. Two-and-one-half hours were allowed for the entire test, including objective and subjective portions. Students typically took thirty to forty minutes for the objective portion so that they had close to two hours left to write their essays.

I was present at one of the marking centers the first of the several days spent scoring the essays, so that I was personally aware of procedures. Instructions to markers asked markers to score the essays analytically, in three separate categories: "Content Development and Organization"; "Sentences (structure, grammar, punctuation)"; and "Words (precision,

vocabulary, level of usage, spelling)." The instructions appear in Appendix 1. Because linguistic and educational research has thus far provided more concrete criteria for evaluating sentences than for the other categories, this study focused on rater response to sentences. The key questions underlying this research were: "What do markers respond to in assigning sentence scores?" and "Does the presence of ESL characteristics make a difference in how the markers respond?"

Markers had to rank each of the three main essay categories on a scale of 1 to 9, 1 being the lowest possible score. Scores of 1 to 3 were called "low"; 4-6 "middle"; 7-9 "high." The three scores were subsequently added to render a total "essay" score for the subjective portion of the EPT. The people employed to mark the essays included people teaching or qualified to teach English in high school, college, university, or as graduate teaching assistants, and some people employed as graders in English departments of public schools. Each essay was marked two times, by different people, each grader ignorant of the other's score. If the difference between the two scores was significant, i.e., if there was enough discrepancy between the scores to affect a student's placement, a "senior marker" adjudicated the scores.

## II. Procedures

Since Chinese is the dominant first language among British Columbia's ESL students, I used native-Chinese-speakers' papers for the E-2 group in the study. My intention was to compare E-1 and E-2 papers at every grade level, 1-9. To select samples I first randomly isolated 2,000 of the total of 18,000 papers. I then had to alter my original plan because the two thousand essays contained very few E-2 papers with sentence scores above 5. I found E-2 papers that were scored 1, 2, 3, 4, and 5 by both

markers, but none given 6, 7, or 8 by both, and none given a 9 by either marker. There were a few papers in the upper-middle to high ranges with the mixed scores of 5/6 and 7/8; I gave these papers "split" scores for my purposes--5.5 and 7.5--to extend the spread of these casually-selected samples as far as possible. There seemed to be more split scores for E-2 papers than for E-1; I believe this impression worth examining in future research as possible evidence of rater insecurity regarding evaluation criteria. Also, I should point out that future research should examine rater responses to a variety of first-language groups, rather than to one, as was the case in this small pilot study.

The difference between a 3 (a high "low" score) and a 4 (a low "middle" score) could mean a difference for the student between an ESL grammar course as opposed to a regular college composition course, and/or a non-credit vs. a credit course. Hence, I became especially interested in marker's differences in this borderline area. Both to get enough samples, and to look more closely at papers over which markers differed, I decided to include split scores, or half steps, through the whole spread of my E-2 samples. This decision gave me eleven E-2 papers, which had the following scores: 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 7.5. I had to complete my E-1 spread with a split score, as well, for I found no papers with two 9's for a top score. My nine E-1 papers, then, ranged from 1 to 8.5. Once the papers had been selected, a research assistant typed these samples and identified the writers with pseudonyms.

I was now ready for the question of what seemed to motivate markers to give the papers the sentence scores they had given them. My first step was to study the relations between the instructions to markers and certain

features of the papers they had marked. I began this step by classifying, in terms that would enable me to count them, those features specifically named in the instructions (Appendix 1). This is how I grouped and classified those sentence features: 1) Fascile use of a variety of sentence structures; 2) Variety in sentence length; 3) Variety in punctuation, including noticeable, discriminating use of the semi-colon; and 4) Errors in sentence structure, grammar, and punctuation that obscured meaning or impaired the flow of ideas. Then, to analyze relations between the sentence scores and these sentence features, I employed the following means (numbers in parentheses indicate columns where this information appears on Table 1).

1. For structural variety and subordination:

   I counted the sentences in the essays and the words, clauses, and T-units of the sentences, and then applied the syntactic maturity indexes developed by Kellog Hunt to analyze syntactic maturity: Words per T-unit (1); words per clause (2); clauses per T-unit (3); and T-units per sentence (4).[6] I also counted the clauses with finite passive verbs, because such clauses have been judged to indicate syntactic maturity to raters in other research on evaluation (5).[7]

2. Sentence length: I counted the words per sentence to find the shortest and longest sentence (6-7) in each paper.

3. Punctuation variety: I counted the number of different uses of punctuation employed, other than the period at the end of the sentence (8), and I checked whether semi-colons were employed and used correctly (9). Punctuation variety can indicate structural variety; besides being considered per se as a feature markers were instructed to evaluate, the relation of this feature to structural variety and subordination (number 1 above) should not be overlooked. Next I counted the number of words in the

essays, because essay length has been found to take priority over sentence structure in scores on writing quality.[8]

4. Errors: I counted seven types of errors in sentence structure, grammar, and punctuation. However, I did not attempt to limit the count to the errors that obscured meaning or impaired flow, as directions stated, because such judgments are too subjective, and I also assumed that the presence of an obvious error would distract a substantial number of markers. The literature on evaluation indicates that errors do affect grades on composition heavily. I counted and classified the errors, as follows: errors in subject and verb agreement (11); other errors in finite verb forms (12); sentence fragments (13); comma splices and run-on sentences (14)--these are typically treated as grammatical and not punctuation errors; unclear or ambiguous pronoun references (15); awkward or illogical constructions (16), where I included misused articles; and punctuation errors (17). I also tabulated the total of all these errors ("sub-total": column 18).

My next step was to count the number of misspelled words, because spelling errors also sometimes influence sentence scores.[9] Following the total misspelled words on Table 1, I indicate the number of different words misspelled in parentheses (columns 19-20).

## III. Results

There were finally twenty-one items to correlate with the sentence scores. When I completed the tabulations, I correlated the relations between the features I had counted with sentence scores for the E-1 and E-2 groups, using the Pearson product-moment statistical procedure. I also correlated scores on the other two principal rating criteria ("Words" and "Content Development and Organization") with the sentence scores. The results showed low correlations ($r < .10$) between most of the sentence

features and the scores the markers had given to sentences, and no significant similarities between the score/sentence feature correlations of the E-1 and E-2 groups.

Because of the small number of samples, correlation coefficients had to be quite high to remain statistically significant. Table 2 distinguishes correlations at the highest .01 level of significance (*); at the high .02 level (V), at the .05 level (X), and at the .10 level (+). The .10 level is not usually considered statistically significant, but this being a preliminary study, the .10 level suggests a correlation worth examining further since it indicates that probably only one time in ten would the relationship occur by chance. Moreover, although there are too few samples here to draw any conclusions, when a high correlation is found among relatively few samples, the probability that such correlation is significant is strengthened by the odds against getting such a coefficient by chance.

The following discussion tells what the Pearson correlations disclosed concerning what correlated significantly with sentence scores on the test.[10]

A. E-1 Correlations

(1) There were no significant correlations between any of the four syntactic maturity indexes and sentence scores;[11] or between passives and scores.

(2) There were no significant correlations between sentence lengths and sentence scores.

(3) There were no significant correlations between punctuation variety or use of the semi-colon and sentence scores.

(4) There was a negative relationship between "awkwardness" and sentence scores of high significance at the .02 level, indicating that

the fewer problems writers had with awkwardness, the higher their scores. Awkwardness, then, is the only feature of the sixteen counted that had a predictably high relationship to sentence scores.

On the other hand, there was an unpredictable positive correlation between scores and sentence fragments at the .10 level. In Table 1, the distribution of fragments among E-1 writers shows us how the samples led to this correlation. Fragments, widely considered to be a major sentence error, did not occur among sentences rated 1, 2, 3, and 4. However, the essays ranked highest on sentences, 8 and 8.5, both contained fragments; the paper given an 8 by both markers had three fragments. Clearly, fragments did not downgrade papers from "high" to "middle" in these cases, even though notes to markers say specifically that the scores of 7, 8, and 9 should go to papers in which "errors, if any, are isolated and relatively minor." No other errors correlated with sentence scores at the .10 level or less, and the correlation of the total sentence errors with sentence scores was not significant.

B. E-2 Correlations

1-2. Again, there were no significant correlations between scores and syntactic maturity indexes, passives, or sentence lengths.

3. Punctuation had an interesting negative correlation: the less flexibility in punctuation, the higher the score, or, conversely, the greater the skill demonstrated in using punctuation, the lower the score, at the .10 level of significance.

4. Among sentence errors, reference problems also had correlated in the opposite way from what could have been anticipated. Reference problems had a positive correlation at a high--.02 level--level of significance.

Thus the more reference problems there were, the higher the scores on sentence quality. Awkwardness did not have a significant correlation as it did in the E-1 group. Comma splices/run-on sentences correlated negatively at the .05 level of significance, and both the total sentence errors (column 18, Table 1) and the total of sentence errors plus spelling errors (column 20), had negative correlations with sentence scores at the .10 level.

The reversals from what instructions lead one to expect raise interesting questions. Do markers not recognize reference problems, or do they not give them much weight? Do the sorts of sentences that raters value tend to contain structures that require pronoun references, at which these ESL students are not very adept? And does the mastery of a variety of other types of structures, the presence of which would be indicated by the skillful punctuation, go unnoticed, or just unrewarded?

Considering the results of the data analysis, common sense suggests that it would be wrong to believe that all the features that correlated significantly with sentence scores were <u>influences</u> on those scores. Only in the cases of awkwardness for E-1 writers and of the various error counts for E-2 writers does influence seem likely. It does appear that errors had a stronger effect on the scores of E-2 writers than on the scores of the E-1 group. Table 1 again helps to make the statistics more meaningful by giving the counts for each individual sample.

C. <u>Correlations of Non-Sentence Features for Both Groups</u>

Essay length correlated at the very significant .01 level for E-1 students; the correlation for E-2 students was not significant. The E-2 student scoring 2 wrote 802 words, a fact that the raters probably

penalized according to instructions. With double the necessary essay length, the writer probably also doubled the sentence errors that he/she would have made in 400 words; 48 is the greatest number of errors among the 20 essays on the chart. The E-1 writer who scored 8 wrote the second highest number of words in the E-1 group, 571. She/he also made more sentence errors than any other E-1 writer, and more spelling errors. It was noted above that error sub-totals and totals correlate at the .10 level of significance for E-2 students, but the correlations on those items are not significant for the E-1 group.

The answer to the question of what most influenced sentence scores might be found in the following explanation: in both groups, E-1 and E-2, the other two essay-scoring categories, "Content Development and Organization" and "Words," correlated with the sentence scores at the highly significant .01 level. It seems reasonable that skill with diction, content and organization, and sentences would often go together. But it might be worth questioning whether the correlation is naturally as strong as the figures suggest--so strong that the higher the score on one, the higher the score on the other, at least ninety-nine percent of the time. Table 1 may contain some hints that there was a tendency to give essays similar scores in all categories, across the board. For example, although spelling is a focal point in scoring the "Words" category, the essay with a score of 8 on "sentences" scored 7 on "words"--still in the "high" zone, despite ten different words misspelled. This paper is the one with the most sentence errors in the E-1 group, as mentioned earlier. Such cases should be studied carefully in future research.

The question of whether markers evaluate E-1 and E-2 writers the same way does seem to deserve further study. Yet the data presented here

indicate that problems may be more fundamental than the matter of getting equity between native and non-native English speakers. Table 1 raises serious questions about the reliability and validity of analytical marking, in general. There is no reason to think that the reliability and validity of EPT scores was not typical of many writing assessments. At the 1983 Conference on College Composition and Communication (Detroit, March 17), Dr. Michael A. McCready, Professor of Education at Louisiana Tech University, reported on a study his university conducted with the National Institute of Education "to determine the 'state of the art' of writing assessment in the nation." The researchers held a conference for people involved in large-scale writing assessment programs throughout the U.S., e.g., consultants and test agency representatives. Key consultants described the three major standard methods of scoring used--holistic, primary trait, and analytical. Participants seemed "very proud of the 'new procedures' which had been developed, but although 'much effort, time, and money had been expended' in the development of scoring procedures, . . . in numerous instances, the assumptions underlying standard scoring procedures were being ignored." Indeed, "although serious concerns were raised by the major consultants, the participants appeared to be very pleased with their respective efforts," and to "feel at liberty to adapt standard procedures to local situations without consideration of underlying assumptions and established procedures" (Italics added). As for the training of readers, "the depth and concentration . . . seemed to vary. Very little comment was made as to how reader agreement was established and maintained."[12]

Among the many questions this pilot study has raised, some of the most obvious are: Do readers have difficulty in distinguishing the

categories when scoring analytically? Are readers too error-oriented when they are supposed to consider a variety of sentence features? Are they even more error-oriented when evaluating papers that contain errors not commonly made by native English speakers? In other words, might errors of native-English speakers be more acceptable because they are familiar? When ESL errors relate more strongly to sentence scores than do non-ESL errors, and sentence scores correlate with content scores at a highly significant level, does the focus on errors in ESL papers place ESL students in double jeopardy?

A recent article concerning ESL compositions, only, indicated the importance of error to ESL teachers who do the marking. In this article, Kyle Perkins of Southern Illinois University reported on his efforts to learn, among other things, what objective measures of composition "discriminate between various levels of attained writing proficiency" and whether a standardized, objective writing test can "discriminate between various levels of attained writing proficiency."[13] Perkins concluded that, for his students at the Center for English as a Second Language at Carbondale, "objective measures which do not take the absence of errors into account are of no use in discriminating among holistic evaluations at one advanced level of proficiency."[14] Measures that count error-free units are very good discriminators. When Perkins compared results from the Test of Standard Written English with holistic and objective measures to see whether the TSWE would discriminate among holistic evaluations, the results were not statistically significant. The TSWE, which evaluates the ability to recognize standard written English, would have placed all twenty-nine of Perkins' students in the most advanced intensive English course at Carbondale into a "bonehead" composition course. In most

places, such general remedial composition courses are not prepared to address very specifically the weaknesses that would place ESL students there, and such lack of relevancy could impede the chance for students to ascend the academic ladder. That markers emphasize error in judging the quality of compositions may be indicated by the correlation Perkins discovered between error-free T-units and holistic evaluations; it was significant at the .005 level. In other studies, errors have also appeared to be a significant influence on scores when the subjectswere heterogeneous, i.e., not isolated according to language background.[15]

The data on the two tables presented here could provoke many other questions, but it is time to return to the major question in the present pilot study, which was whether markers of large-scale essay tests appear to respond the same way toward essays written by ESL and non-ESL students. The answer, in the terms of this limited investigation, is that they do not. Yet it appears that to ensure equity in grading essay tests, researchers will first have to define, more specifically, how to judge what is "good" as well as "bad," and how much weight to give these things. Then they need to develop means of retraining the teachers who grade the papers to achieve more uniform evaluations.

## Notes

[1]Rosentene B. Purnell, "A Survey of the Testing of Writing Proficiency in College: A Progress Report," College Composition and Communication, 33 (December 1982), 410.

[2]Purnell, p. 410.

[3]Dorothy Waggoner, "Non-English Language Background Persons: Three U.S. Surveys," TESOL Quarterly, 12 (September 1978), 247.

[4]"Immigration Policy Partly Reformed: Senate Approves Immigration Reform Bill," America, 147 (September 4-11), 103.

[5]"Immigration Policy," p. 103.

[6]Syntactic Maturity in Schoolchildren and Adults, monographs of the Society for Research in Child Development, Serial No. 134, Vol. 35, No. 1, February, 1970. Hunt developed the concept of the T-unit to simplify counting irregular sentence structures. Hunt found words per T-unit (a "main clause plus any subordinate clause or non-clausal structure . . . attached to or embedded in it") to be the best index of syntactic maturity in school children and adults. Words per clause was second, clauses per T-unit, third. Clauses per T-unit provide a subordination index. T-units per grammatically complete sentence index coordination.

[7]In "Syntactic Skill and ESL Writing Quality," a paper presented at the conference on the Canadian Council of Teachers of English in Ottawa 12 May 1979, Patrick Kameen reported that, on ESL compositions that he compared, markers appeared to react positively to longer T-units, longer clauses, and passive constructions.

[8]Ellen W. Nold and Sarah W. Freedman, for example, found that composition length is significant as a variable in determining quality ratings ("An Analysis of Readers' Responses to Essays," Research in the Teaching of English, Vol. 11 (1977), 164-174). Standard developmental measures like words-per-T-unit were not useful in predicting perceptions of writing quality.

[9]I thank Professors David Crough and Faith Gilroy, of the Psychology Department of Loyola College in Maryland, for answering my questions about statistics.

[10]Murray F. Stewart and C.H. Grobe, "Syntactic Maturity, Mechanics of Writing and Teachers' Quality Ratings," Research in the Teaching of English, 13 (1979), 207-215.

[11]This result corresponds with the results of other studies examining the relations between scores on writing quality and such objective measures of syntactic maturity. Besides Nold and Freedman, above, see, for example, Lester Faigley, "Names in Search of a Concept: Maturity, Fluency, Complexity and Growth in Written Syntax," College Composition and Communication, 31 (October 1980), 291-310. These studies, however, correlated essay quality scores, not sentence quality scores, with the objective measures of syntactic maturity.

[12]The gist of Professor McCready's paper appears in Notes From The National Testing Network in Writing, January, 1983, pp. 18 and 22.

[13]Kyle Perkins, "Using Objective Measures of Attained Writing Proficiency to Discriminate Among Holistic Evaluations," TESOL Quarterly, 14 (March 1980), 59.

[14]Perkins, p. 64.

[15]One study by Winifred Hale Harris found that teachers look at sentence structure "primarily in terms of technical correctness rather than looking for variety in either structural elements or kinds of sentence patterns"; "Teacher Response to Student Writing: A Study of the Response Patterns of High School English Teachers to Determine the Basis of Teacher Judgment of Student Writing," Research in the Teaching of English, 11 (1977), 176.

TABLE 1

| SCORE | I MATURITY INDEXES | | | | II RATING INFLUENCES | | | | | III COUNTABLE ERRORS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 W/T | 2 W/C | 3 C/T | 4 T/S | 5 Pass. | 6-7 S.Length | 8 P.Var. | 9 ; | 10 No. Words | 11 Agr. | 12 Verb | 13 Frag | 14 CS/RS | 15 Ref. | 16 Awk | 17 P | 18 Sub-total | 19-20 Spelling | 21 Total |
| **ENGLISH AS A FIRST LANGUAGE** | | | | | | | | | | | | | | | | | | | |
| 1 | 16.83 | 10.82 | 1.56 | 1.20 | 5 | 5-36 | 4 | | 303 | 0 | 1 | 0 | 3 | 3 | 9 | 6 | 22 | 13(12) | 34 |
| 2 | 15.25 | 7.79 | 1.96 | 1.60 | 9 | 11-47 | 9 | | 366 | 0 | 1 | 0 | 8 | 2 | 7 | 1 | 19 | 10(7) | 26 |
| 3 | 8.92 | 7.25 | 1.23 | 1.09 | 0 | 4-15 | 3 | | 237 | 0 | 3 | 0 | 1 | 1 | 4 | 0 | 9 | 7(4) | 13 |
| 4 | 19.17 | 8.21 | 2.33 | 1.00 | 3 | 11-34 | 4 | | 345 | 0 | 1 | 0 | 0 | 0 | 4 | 5 | 10 | 2(1) | 11 |
| 5 | 14.99 | 10.82 | 1.38 | 1.14 | 4 | 3-28 | 7 | | 357 | 0 | 2 | 1 | 0 | 0 | 2 | 3 | 8 | 3 | 11 |
| 6 | 16.09 | 11.92 | 1.36 | 1.19 | 13 | 12-39 | 6 | | 402 | 0 | 2 | 0 | 2 | 2 | 2 | 1 | 9 | 4(3) | 12 |
| 7 | 22.95 | 11.21 | 2.04 | 1.09 | 5 | 9-52 | 5 | ✓ | 597 | 0 | 1 | 0 | 0 | 0 | 2 | 6 | 9 | 2 | 11 |
| 8 | 15.86 | 10.20 | 1.56 | 1.16 | 3 | 3-42 | 8 | | 571 | 0 | 0 | 3 | 4 | 3 | 6 | 7 | 23 | 14(10) | 33 |
| 8.5 | 17.55 | 9.67 | 1.82 | 1.29 | 1 | 10-38 | 5 | | 474 | 1 | 0 | 1 | 2 | 2 | 1 | 9 | 16 | 2 | 18 |
| **ENGLISH AS A SECOND LANGUAGE** | | | | | | | | | | | | | | | | | | | |
| 1 | 10.27 | 7.93 | 1.32 | 2.10 | 0 | 12-43 | 9 | | 452 | 1 | 14 | 0 | 12 | 0 | 7 | 2 | 36 | 4(1) | 37 |
| 1.5 | 13.15 | 7.35 | 1.58 | 1.03 | 3 | 5-25 | 7 | | 434 | 2 | 9 | 4 | 1 | 2 | 18 | 2 | 38 | 4 | 42 |
| 2 | 12.94 | 7.57 | 1.71 | 1.22 | 0 | 2-35 | 10 | ✓ | 802 | 1 | 18 | 4 | 9 | 0 | 14 | 2 | 48 | 2 | 50 |
| 2.5 | 17.52 | 7.58 | 2.31 | 1.07 | 1 | 9-29 | 7 | | 508 | 2 | 1 | 0 | 0 | 1 | 6 | 1 | 11 | 1 | 12 |
| 3 | 17.56 | 8.52 | 2.06 | 1.78 | 0 | 10-89 | 10 | ✓ | 281 | 3 | 0 | 0 | 7 | 1 | 6 | 2 | 19 | 15(5) | 24 |
| 3.5 | 17.90 | 8.70 | 2.00 | 1.67 | 1 | 17-35 | 5 | | 261 | 1 | 1 | 0 | 1 | 1 | 3 | 4 | 11 | 2 | 13 |
| 4 | 13.00 | 8.13 | 1.73 | 1.11 | 16 | 4-25 | 6 | | 520 | 4 | 9 | 2 | 0 | 3 | 7 | 1 | 26 | 4 | 30 |
| 4.5 | 13.48 | 6.74 | 2.00 | 1.06 | 1 | 3-28 | 5 | ✓ | 499 | 0 | 9 | 0 | 0 | 0 | 4 | 1 | 14 | 4 | 18 |
| 5 | 12.48 | 7.87 | 1.59 | 1.12 | 4 | 2-29 | 7 | ✓ | 362 | 1 | 0 | 0 | 1 | 2 | 2 | 2 | 8 | 0 | 8 |
| 5.5 | 14.25 | 9.12 | 1.56 | 1.00 | 8 | 6-43 | 6 | | 456 | 4 | 4 | 1 | 0 | 4 | 5 | 3 | 21 | 0 | 21 |
| 7.5 | 15.74 | 8.85 | 1.78 | 1.42 | 8 | 6-35 | 7 | ✓ | 425 | 2 | 2 | 0 | 0 | 5 | 7 | 3 | 19 | 0 | 19 |

Table 2

Correlation Coefficients

of

Sentence Scores and Writing Features

| | | E-1 | E-2 |
|---|---|---|---|
| Sentence-related Features | W / T | .3662 | .2140 |
| | W / C | .4097 | .2734 |
| | C / T | .0550 | .1004 |
| | T / S | -.2325 | -.3115 |
| | Pass. | -.1921 | .4988 |
| | Short. S. | .3673 | -.2131 |
| | Long. S. | .3082 | -.0994 |
| | P. Var. | .2677 | -.5436 + |
| | ; | .2908 | .4011 |
| | Agr | .5030 | .1647 |
| | Verb | -.4555 | -.4936 |
| | Frag. | .5943 + | -.4128 |
| | CS / RS | -.2896 | -.6248 X |
| | Ref. | -.0487 | .7182 ✓ |
| | Awk. | -.6855 ✓ | -.5106 |
| | P | .5393 | .2614 |
| | Total Sentence Errors | -.0499 | -.5456 + |
| Extraneous Features — Spelling-related | Spelling Errors -- Total | -.3661 | -.3383 |
| | Spelling Errors -- No. of Words | -.3828 | -.3613 |
| | Total Errors | -.1915 | -.5951 + |
| Extraneous Features — Essay-related | Essay Length (No. of Words) | .8022 * | -.2117 |
| | Words (Diction) | .9881 * | .8566 * |
| | Content Development and Organization | .9814 * | .8021 * |

* p .01

✓ p .02

X p .05

\+ p .10

# Appendix 1

## ENGLISH PLACEMENT TEST

### NOTES TO MARKERS

In order to achieve a measure of uniformity among a large group of markers, it is necessary to evaluate each essay according to a set of categories. There will be some overlap between categories, and between the classifications within each category. Moreover, an essay may belong to one of the classifications without necessarily displaying all of the symptoms described under that classification.

Some of the more common difficulties of classification are as follows:

(1) Essays that are too short or too long. The instructions call for an essay of 300-500 words. Essays which fall outside the limits by more than 100 words in either direction should be penalized for "Content Development and Organization." Those essays which are shorter than required should be more severely penalized by being ranked in the Low classification.

(2) Narrative. The expository mode is required. While an expository essay can contain some narrative development, an over-dependence on narrative might be construed as a failure to properly address the topic. Unless a thesis that has a direct bearing on the topic is clearly implied in such an essay, the essay should receive no more than a Low ranking for "Content Development and Organization."

(3) Short paragraphs. One of the most common faults of the student writer is the failure to properly develop, expand, and support his main points. The result: a series of short, one-, two-, and three-sentence paragraphs. This lack of proper development should be reflected in the score for "Content Development and Organization," and should be distinguished from the paragraph fragment, which occurs when an idea, although adequately developed, is unnecessarily fragmented into more than one paragraph. This structural fault should be reflected in the score for "Content Development and Organization."

Reminders

We should not penalize an obviously sophisticated writer who is deliberately "violating" a rule, or varying a pattern for effect; e.g. a slang expression consciously introduced, a sentence fragment used as a virtual sentence, a one-sentence paragraph employed to emphasize an important point, etc.

It is important to ensure that non-native writers are not penalized in all three criteria areas for errors in sentences and words only -- unless the errors are serious enough to obscure meaning or impair the flow of ideas.

### CRITERIA FOR SCORING ESSAYS

1. CONTENT DEVELOPMENT AND ORGANIZATION

High. (Ranked 9, 8, or 7). The paper has an overall plan, point of view, and sense of direction. The thesis and the main supporting points are original and interesting, and are developed fully in a logical, persuasive, thought-provoking manner. Each paragraph is unified around a clear focus. The transition within and between the paragraphs is clear and smooth (without being mechanical), allowing the reader to move from point to point with a sense of coherence and momentum.

Middle. (Ranked 6, 5, or 4). Content and organization are adequate but pedestrian. The essay has a main idea, although it is not clearly stated or kept in focus throughout. The content is skimpy -- often short, one- and two-sentence paragraphs restricted to generalizations and commonplaces. The arguments are too one-sided. There is no clear organization, or it is too mechanical. Transition is lacking, or it is restricted to the simplest devices. The order of points is adequate but nothing more.

Low. (Ranked 3, 2, or 1). No central idea governs. Obvious points are repeated in slightly different words. The essay pays lip service to the topic, but actually avoids it. (Essays that offer unreasonably one-sided arguments should also be placed in this category.) Points come in haphazard order, without coherence or a sense of direction. The paragraphing contributes little or nothing to the flow of ideas. Transitional devices are lacking or extremely mechanical.

BEST COPY AVAILABLE

2. SENTENCES (STRUCTURE, GRAMMAR, PUNCTUATION)

High. (Ranked 9, 8, or 7). The sentences are clear and demonstrate a variety of length and structure. There is clear evidence that the writer is able to use subordinate structures, and is not limited to simple and compound sentences. Errors, if any, are isolated and relatively minor.

Middle. (Ranked 6, 5, or 4). The sentences lack variety in length and structure. The writer is able to construct familiar sentence patterns, but is prone to error and avoids sophisticated structures, such as parallelism, subordination, etc. There is reasonable control of the comma, but semi-colons and other marks of internal punctuation are absent or used incorrectly or indiscriminately.

Low. (Ranked 3, 2, or 1). There are errors in sentence structure, grammar, or punctuation such as to obscure the meaning or seriously impair the flow of ideas.

3. WORDS (PRECISION, VOCABULARY, LEVEL OF USAGE, SPELLING)

High. (Ranked 9, 8, or 7). Words are used precisely and imaginatively. The vocabulary appears to be wide-ranging. Wordiness and pretentious diction are absent. The diction is specific and concrete where appropriate -- not restricted to the abstract and general. The level of usage is consistent, neither too formal nor too informal (except for deliberate effect). Misspellings, if any, are few, and not "glaring".

Middle. (Ranked 6, 5, or 4). The phrasing is standardized, ordinary -- sometimes hackneyed, never fresh. There is evidence of imprecision and wordiness. There are inconsistencies in the level of usage, especially in the direction of overly informal expressions. Some jargon or pretentious diction is evident. Spelling is perhaps a noticeable problem, with some errors in regular words as well as in more difficult words.

Low. (Ranked 3, 2, or 1). Imprecise and wrong words abound. The phrasing is vague, general, wordy. The level of usage is frequently too informal. Spelling is a definite problem. An essay written in an overly simplistic vocabulary may also get a Low rating, even if no word is clearly wrong. Similarly, excessive wordiness or gobbledygook deserves a Low rating.

BEST COPY AVAILABLE